

# PATENT ABSTRACTS OF JAPAN

(11)Publication number : 03-014184

(43)Date of publication of application : 22.01.1991

(51)Int.Cl.

G06F 15/62

G06F 15/40

(21)Application number : 01-148463

(71)Applicant : FUJI XEROX CO LTD

(22)Date of filing : 13.06.1989

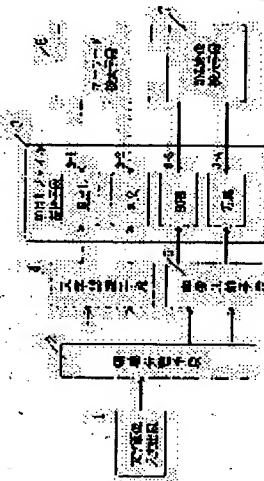
(72)Inventor : HAYASHI YUKIO

## (54) DOCUMENT IMAGE REARRANGEMENT FILING DEVICE

### (57)Abstract:

**PURPOSE:** To efficiently execute a retrieval by extracting and separating automatically areas of different attributes such as a heading, a text, a chart, a photograph, etc., in a document image and accumulating them at every item so that the retrieval can be executed by a suitable retrieving method at every area of the attributes.

**CONSTITUTION:** An input document image inputted by scanning a newspaper, etc., by a document image input means 1 is separated into each attribute area of a heading 3-1, a text 3-2, a chart 3-3, and a photograph 3-4. Subsequently, as for the areas of the heading 3-1 and the text 3-2, a character recognition processing is performed and they are converted to code data, and as for the areas of the chart 3-3 and the photograph 3-4, an image compression processing is performed, and thereafter, they are brought to filing to each separate attribute item. In the retrieving means 6, or the retrieving means 7, the retrieval can be executed by selecting one of a key word retrieval to the heading 3-1 or the text 3-2, and a retrieval of a similar image to the chart 3-3 and the photograph 3-4. In such a way, the retrieval/display time for a document image is shortened, and also, by rearranging the document image, the data compression can be executed.



## LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2000 Japanese Patent Office

## SPECIFICATION

### 1. TITLE OF THE INVENTION

DOCUMENT IMAGE RE-ARRANGEMENT FILING APPARATUS

### 5 2. WHAT IS CLAIMED IS:

A document image re-arrangement filing apparatus comprising:

document image input means for inputting document image;  
region dividing means for automatically extracting from  
10 the document image input by said document image input means  
different attribute regions such as a headline region, a text  
region, a drawing region and a photograph region, and dividing  
them;

itemized file storing means for accumulating and storing  
15 items such as the headline, the text region, the drawing region  
and the photograph region, extracted and divided by the region  
dividing means

retrieving means for retrieving data of the items stored  
in the itemized file storing means, by item;

20 document image re-arrangement means for retrieving  
desired document image by retrieval of the retrieving means,  
and re-arranging the data in the items to output the entire  
document image.

### 3. DETAILED DESCRIPTION OF THE INVENTION

25 [FIELD OF THE INVENTION]

The present invention relates to a document image re-arrangement filing apparatus capable of automatically extracting images in regions consisting of various attributes such as a headline, a text, a drawing, a photograph and the like, to store item by item, re-arranging each of the entire item in image such as a scrap of a newspaper.

[PRIOR ART]

It is proposed now for an apparatus that files document images such as a scrap of a newspaper to employ a system for automatically extracting the complicated layout shapes according to each attribute (a headline, a text, a drawing, a photograph) region of the document image, and re-arranging the image data compactly.

For example, there is proposed a database editing and controlling system that extracts image of a newspaper that is input by an image input apparatus, as a pattern with use of a man-machine interface, stores the image data after structuring the data by structurally rewriting the data to reduce the data amount, and re-edits the data in use in accordance with the object (The Institute of Electronics, Information and Communication Engineers GIHOU PRL84-101, pp.65-72).

Similarly, an automatic layout system capable of automatically extracting the document structure such as a hierarchical structure based on the concepts such as a title, a chapter and a paragraph, and a reference structure between

a text and drawings by using format information of the document or keywords, and executing formatting or layout of the document with use of a format based on the document structure and knowledge about layout technique (JOHOSHORI ASSOCIATION "DOCUMENT  
5 PROCESSING AND HUMAN INTERFACE", 20-3, September 8, 1988).  
[OBJECT TO BE SOLVED BY THE INVENTION]

In order to obtain the desired document image with use of the above-mentioned techniques, however, the entire document image has to be retrieved and checked, which is a so troublesome  
10 job requiring a long period of time. Further, the image document contains various data such as a headline, a text, a drawing, a photograph and the like in a mixed manner, so that in some cases, it turns out wasteful from a viewpoint of data storage system.

15 The present invention intends to provide a document image re-arrangement filing apparatus capable of reducing a period of time required for retrieving and displaying the document image, and compressing data by re-arranging the document image data to remove redundant space in the document.

## 20 [MEANS OF SOLVING THE PROBLEMS]

As shown in Fig. 1, the document image re-arrangement filing apparatus according to the present invention comprises: document image input means 1 for inputting by scanning document image such as a scrap of a newspaper; region dividing means 2  
25 for automatically extracting from the document image input by

the document image input means 1 different attribute regions such as a headline region(3-1), a text region (3-2), a drawing region (3-3), a photograph region (3-4) and the like, and dividing them; itemized file storing means 3 for accumulating and storing

5 items such as the headline, the text region, the drawing region, and the photograph region, extracted and divided by the region dividing means 2; retrieving means (6, 7) for retrieving data of the items stored in the itemized file storing means 3, by item; document image re-arrangement means 8 for retrieving.  
10 desired document image by retrieval of the retrieving means (6, 7), and re-arranging the data in the items to output the entire document image.

[EFFECT]

The input document image data input by scanning a newspaper  
15 or the like with use of a document image input means 1 is divided into each attribute region such as a headline 3-1, a text 3-2, a drawing 3-3, and a photograph 3-4 by a region dividing means 2, and then positional data of the rectangular regions enclosing the respective regions is then extracted: coordinates (x, y)  
20 of an upper left corner of the rectangular region; a width w; and a height h. At this time, it is determined whether the document is written vertically or horizontally. Subsequently, the attribute regions the headline 3-1 and the text 3-2 are subjected to the character recognition process to be converted  
25 into code data, and the attribute regions the drawing 3-3 and

the photograph 3-4 are subjected to the image compressing process, and the image data is filed into the corresponding items. The retrieving is executed by either a retrieving means 6 or a retrieving means 7: keyword retrieving (substring matching for  
5 registered portion key) executed for the headline 3-1 or the text 3-2 by retrieving means 6; and analogous image retrieving executed for the drawing 3-3 and the photograph 3-4 by retrieving means 7. During retrieving, only one of the image data among the headline 3-1, the text 3-2, the drawing 3-3 and the photograph  
10 3-4, which is now retrieved is displayed. When the retrieving has finished, a re-arrangement processing means 8 executes the re-arrangement process for the entire document and displayed. The re-arrangement processing means 8 executes repeatedly the re-arrangement of the retrieved data to arrange the retrieved  
15 data from the upper right towards left in order of the headline 3-1, the text 3-2, the drawing 3-3 and the photograph 3-4, and a line is filled with the data, arrange the data to the other paragraph (in the horizontally written document; from the upper left toward right).

20 Figs. 2 and 3 respectively show the division result of the image data into the attribute regions and the re-arrangement result.

#### [EMBODIMENTS]

The technical idea of the present invention will be  
25 described below while referring to Fig. 1.

In this drawing, the document image input means 1 scans image document such as a scrap of a newspaper and input the image data, and the region dividing means 2 automatically extracts the attributes of the image data from the input document image, for example, the headline 3-1, the text 3-2, the drawing 3-3 and the photograph 3-4.. Among the document image extracted with each attribute items, the headline 3-1 and the text 3-2 are recognized as characters by a character recognition means 4 to be converted into character codes, and then stored in an itemized file storing means 3. When the extracted document image is image data, the image data compressing means 5 compresses the data to be stored in the itemized file storing means 3. If the stored image needs to be displayed, the data can be retrieved with use of the keyword retrieving means 6 for retrieving character data from the headline 3-1 and the text 3-2 as by item and the analogous image retrieving means 7 for retrieving image data from the drawing 3-3 and the photograph 3-4. The document image re-arrangement means 8 displays the final retrieving result obtained by these retrieving means.

The regions can be arbitrarily recognized and divided according to the difference of the attributes of the regions; area, density, the number of edges, or the like of a text region, a headline region, a drawing region, a photograph region and the like.

Fig. 4 is a block diagram showing the constitution example

of the region dividing means 2. The constitution shown in the diagram is based on the technical inventions disclosed in Japanese Patent Application No. SHO 63-252920, Japanese Patent Application No. SHO 63-291095, and Japanese Patent Application  
5 No. HEI 1-87039 (inventor of all the applications: Yutaka NAKAMURA) by the applicant of the present application, comprises:

binarizing means 201; rectangularized region image storing memory 202 for storing the rectangularized image; rectangularizing means 203 for rectangularizing image;  
10 binarized image memory 204 for storing binarized image; isolated point removing means 205 for removing isolated points from the binarized image; outline extracting means 206 for extracting outline of the image after removing the isolated points; outline image memory 207 for storing the extracted outline image; and  
15 image region recognizing means for recognizing image region on the basis of the data stored in the memories 202, 204, and 207.

The image region recognizing means has rectangular region address generating section 208, black pixel calculating section 209 and 210, text character recognition section 211, rectangular area  
20 calculating section 212, comparing section 213, area comparator (headline/drawing distinguishing section) 214, comparator (photograph/character-linear image distinguishing section) 215 and the like.

The image is input by a scanner or the like in a multivalued  
25 form. The input image data is binarized by the binarizing means



201. The binarized image data is subjected to the  
rectancularizing process executed by the rectancularizing means  
203, wherein the black pixels are connected and the connected  
black pixels are enclosed by a rectangular. The obtained  
5 rectangular region is stored in the rectancularized region image  
storing memory 202, and used as an image recognizing unit.

After the binarizing process, the isolated points of the  
binarized image data are removed by the isolated point removing  
means 205, and the outline, i.e., the edge of the image is then  
10. extracted by the outline extracting means 206, and stored in  
the outline image memory 207.

The recognition of the region is performed as follows:  
at first, the text character recognition section 211 recognizes  
the text area by inferring the character height of the text on  
15 the basis of the distribution of the heights of the rectangular  
portions in unit of the region enclosed by the rectangular  
obtained by the rectancularizing means 203, and divides the text  
areas. Next, the headline regions, the drawing regions, and  
the photograph regions other than the text regions are recognized.  
20 on the basis of the area, the density and the edge of the rectangle.  
That is, the rectangular region address generating section 208  
generates the addresses of the pixels in the rectangular regions  
other than the text region on the basis of the rectancularized  
image data read from the rectancularized region image storing  
25 memory 202. The pixels in the rectangular region are read from

the binarized image memory 204 and the outline image memory 207 in accordance with the generated addresses. The number of the black pixels in the read pixels are respectively calculated by the black pixel calculating sections 209 and 210. The calculated  
5 number of the black pixels in the binarized image in the rectangular portion is compared with the number of the black pixels in the outlined image by the comparing section 213 to obtain the ratio thereof. The comparator 215 compares the ratio obtained by the comparing section 213 with a predetermined  
10 threshold value. When the ratio is larger than the threshold value, the image is determined as a character-linear image region, and when the ratio is smaller than the threshold value, the image is determined as a photograph region.

When the image is determined as a character-linear image  
15 region, the image is further determined on the basis of the area of the rectangle whether it is the headline region or the drawing region. In order to distinguish the image regions in this way, the area of the rectangular region is calculated by the  
20 rectangular area calculating section 212 on the basis of the rectangularized region image read from the rectangularized region image storing memory 202. Subsequently, the area comparator (headline/drawing distinguishing section) 214 compares the calculated area of the rectangular region with a  
25 smaller than the threshold value, the rectangular region is

determined (recognized) as the headline region, and when the calculated value is larger than the threshold value, the rectangular region is determined as the drawing region.

On the basis of the determination (recognition signal),  
5 the image signal of each region is divided and extracted from the input image stored in the image memory or the like, and subjected to necessary processes such as character recognition, code compression, and addition of the attribute information for retrieving, then registered in the itemized file storing section  
10 3. That is, the scan-in input image is divided into the attribute region such as the headline region 3-1, the text region 3-2, the drawing region 3-3, and the photograph region 3-4, and the positional data of the rectangular regions is then extracted: coordinates (x, y) of an upper left corner of the rectangular  
15 region; a width w; and a height h. At this time, it is determined whether the document is written vertically or horizontally. Subsequently, the headline region and the text region are subjected to the character recognition process by the character recognition means 4 so as to be converted into code data. The  
20 drawing region and the photograph region are subjected to the image compressing process by the image data compressing means 5. The itemized file storing means 3 files the processed data of the attribute regions by item.

The registered image in the above-described manner is  
25 performed by the keyword retrieving means 6 in accordance with

the keyword retrieving (substring matching for registered portion key) to retrieve the required image. The retrieval of the drawing or photograph is executed by the analogous image retrieving means 7.

5       An example of the analogous image retrieving executed for the drawing 3-3 and the photograph 3-4 will then be described in detail with reference to Figs. 5 and 9. The example is based on the description disclosed in Japanese Patent Application No.

HEI 1-113829 (inventor: MAKOTO HIROSE) by the applicant of the

10       present application, which is advantageous in that the required image data can be retrieved at a high speed.

Fig. 5 is a block diagram showing the schematic constitution of an embodiment the image retrieving apparatus capable of performing the above-mentioned analogous image  
15       retrieving according to the present invention. The image retrieving apparatus comprises: an image input device 25 used by a user or an administrator for inputting image; a display device 26 for displaying image or the other information to the user or the administrator; a calculation processor 27 for  
20       calculating the degree of the agreement of the images and controlling the order of displaying the retrieved image or entire apparatus; an image output device 28 for outputting the retrieved image in a form of printing; a control instruction input device  
25       29 used by the user to input the tolerance of the image agreement determination; and a memory device 30 for storing image and

information.

In order to execute the analogous image retrieving effectively, the drawing and photograph are processed to add the characteristic information in registering in the itemized  
5 file storing section 3.

The registering operation will be described at first. Fig.

6 shows an example of the structure of the two-dimensional binarized image information sent from the image input device

25. The image information is displayed as image by the display

10 device 26. The whole image information is binarized data of

$M \times N$  wherein 0 expresses white pixel, and 1 expresses black

pixel. The calculation processor 27 prepares character

information 32 constituted in an architecture shown in Fig. 7

on the basis of character quantity obtained from the quantization

15 of the number of value 1 every small region of  $K \times L$  in the respective

regions. The character information 32 and the  $M \times N$  image

information 33 are stored in the memory device 30. In this manner,

as shown in Fig. 5, a character information group 34 and the

image data group 35 of a plenty of images are stored in the memory

20 device 30. The head image number 31 is an arbitrary number given

for the calculation processor 27 to recognize a plurality of

image data.

The retrieving process of the image data will then be described next with reference to the operation process flow chart

25 shown in Fig. 9. The number 1 of the images desired by users

is input into the control input device 29 as the tolerance of the image agreement determination (1).

Upon receiving the image as the retrieving condition, the calculation processor 27 prepares character information  $a_{K1}$  of the image as the retrieving condition, in the same manner as the registration (2).

After setting  $j$  as 1, the value of a register  $D_I$  to which  $I_{th}$  value is input in order of the degree of disagreement from the smaller one is set as  $K \times L$  (3).

10 The character quantity  $b_{JK1}$  of the image number  $j$  is calculated (4). As shown in Fig. 8, In the case where the character quantity is calculated in registering the image and registered together with the image as the index of the image, the registered value is read from the memory device 30.

15 The calculation processor 27 calculates the degree of the disagreement  $D_J$  between the  $j_{th}$  image and the image to be retrieved serially from the image number 1 (5).

In every time when the value  $D_J$  is obtained, the value  $D_I$  as the  $I_{th}$  value in order of the degree of the disagreement from the smaller one is compared with the value  $D_J$  (6).

20 When  $D_J > D_I$ , the desired image is not present in the group to the  $I_{th}$  image, and thus a process on the next image is executed. At this time, the value  $j$  is add with 1 (7).

When the result of the comparison is  $D_J \leq D_I$ ,  $D_I$  is replaced with  $D_J$  (7). The number of images from the image having the

smallest disagreement value to the  $I_{th}$  image are stored in the memory (8). A process on the next image is then executed and the value  $j$  is add with 1 (9).

Finally, after calculating the degree of the disagreements  
5 of all the stored images, the image having the smaller disagreement degree is displayed in the display device 26.

The calculation of the degree of the disagreements between the character information of the images in the character information group 34 and the character information of the image  
10 to be retrieved will be described below. The character information comprises the character quantity expressed by the image number 31 and the matrix having  $K$  rows and  $L$  columns. The character quantity of the image to be retrieved is expressed as  $a_{k1}$  ( $k = 1$  to  $K$ ,  $1 = 1$  to  $L$ ), and the character quantity of  
15 the  $j_{th}$  image in the character information group 14 to be retrieved is expressed as  $b_{jkL}$  ( $k = 1$  to  $K$ ,  $1 = 1$  to  $L$ ). Assuming the number of 1 in each of the small regions is quantized as binary value, the degree of the disagreements between the image to be retrieved and the  $j_{th}$  image is obtained by the calculation expressed by  
20 the following formula:

$$D_j = \sum \sum (a_{k1} \oplus b_{jkL})$$

In other words, the total sum of the exclusive OR of the corresponding elements is defined as the disagreement. The symbol  $\oplus$  denotes the exclusive OR.

25 Assuming the number of 1 in each of the small regions is

quantized as multivalued, more specifically,  $a_{kL}$  and  $b_{jKL}$  are multivalued, the degree of disagreements between the image to be a condition and the  $j_{th}$  image is obtained by the calculation expressed by the following formula:

5 
$$D_J = \sum \sum | (a_{kL} - b_{jKL}) |$$

Prior to the output of the image as the result of the retrieval, the re-arrangement of the entire document comprising the headline 3-1, the text 3-2, the drawing 3-3 and the photograph 3-4 is then performed. Figs. 10 and 11 shows the process-flow of the re-arrangement. The processes shown in Figs. 10 and 11 are connected at (a) and (b).

At first, the headline storing position in the itemized file storing means 3, which is assigned to the retrieved document image is searched to determine whether or not the headline 3-1 is present. (11).

When the headline 3-1 is present, it is determined whether the headline 3-1 is written vertically or horizontally (12). When the headline 3-1 is written vertically, the read headline is re-arranged in accordance with the vertical re-arrangement procedure (13). When the headline 3-1 is written horizontally, the read headline is re-arranged in accordance with the horizontal re-arrangement procedure (14).

In spite of the fact that the headline 3-1 is written vertically arranged (13) or horizontally arranged (14), the headline storing position is further searched to determine



whether or not the other headline is present (11).

When the other headline is not present in the storing means, the re-arrangement has finished and the re-arrangement of the character lines of the text will be executed next. At this time,  
5 the next row is set as the upper right end coordinate of the character lines (15).

The storing means is searched to determine whether or not the other character line is present (16). When the character line is present, the right edge space present on the right side  
10 of the character line is compared with a predetermined margin provided to the left side (17). When the left margin is smaller than the right space, the re-arrangement of the character line is executed (18). The re-arrangement of the character line is executed such that the position subtracted by the width of a  
15 character from the right end of the present character line is determined as the right end of the next character line (19).

When the space on the left side of the character line becomes smaller than the predetermined left margin in the step (17), the paragraph is changed and the right end space of the character  
20 line is set again (20). The re-arrangement of the character line is performed in the same manner as in the step (18) (21).

When the other character line to be arranged is not present in the storing means in step (16), the re-arrangement of the drawing and photograph will be executed next. At this time,  
25 the upper right end coordinate of the drawing and the photograph

(22).

When the drawing or the photograph is present in the document image in step (23), the right edge space present on the right side of the drawing or the photograph is compared with  
5 a predetermined margin provided to the left side (24). When the left margin is smaller than the right space, the re-arrangement of the drawing and the photograph is executed (25). The re-arrangement of the drawing and the photograph is executed such that the position subtracted by the width of the  
10 drawing and the photograph from the right end of the present drawing or the photograph is determined as the right end of the next drawing or photograph (26).

In the case that the predetermined left margin becomes bigger in step (24), the paragraph is changed and the right end  
15 space of the character line is set again (27). The re-arrangement of the drawing or the photograph is performed in the same manner as in the step (25) (28). When it is determined that no more drawing or photograph is present in the storing means, the re-arrangement process has finished (20).

## 20 [ADVANTAGES OF THE INVENTION]

According to the present invention, the different attribute regions such as headline region, text region, drawing region, photograph region or the like in a document image can be automatically extracted from the document image, divided and  
25 classified by item (attribute region) to store them. In addition,

in retrieving the stored image data, the image data can be retrieved in accordance with the process suitable for the attribute region to be retrieved. Further, during the retrieval, only the attribute region now retrieved can be displayed. With  
5 such a retrieving method, the retrieval can be performed so effectively.

Further, the retrieved document image is re-arranged by classifying and collecting them by item, the redundant space in the document can be removed and data can be compressed, and  
10 the period of time required for retrieving can be reduced.

#### 4. BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram showing the constitution of the document image re-arrangement filing apparatus according to the present invention; Figs. 2 and 3 show the modes of filing the  
15 respective attribute regions abstracted by the region dividing;

Fig. 4 is a block diagram showing the constitution of the region dividing means; Fig. 5 is a block diagram showing the constitution of the image retrieving apparatus; Fig. 6 shows the structure of the two-dimensional binarized image data; Fig. 7 shows the  
20 architecture of the character information; Fig. 8 shows the character information group and the image information group; Fig. 9 is a flow chart of the retrieving process; Figs. 10 and 11 are flow charts of the re-arrangement process by the re-arrangement means 8 shown in Fig. 1.

- 1 ... document image input means
- 2 ... region dividing means
- 3 ... itemized file storing means
- 4 ... character recognition means
- 5 5 ... image data compressing means
- 6 ... keyword retrieving means
- 7 ... analogous image retrieving means
- 8 ... document image re-arrangement means

Applicant FUJI XEROX CO. LTD.

10 Attorney SHOUICHI IWAKAMI

Drawing

Fig. 1

①headline

②photograph

5 ③drawing

text (本文)

Fig. 2

②document image

Fig. 3

10 ②headline

Fig. 4

①inputting an image

②binarizing means 201

③rectangularized region image storing memory 202

15 ④binarized image memory 204

⑤rectangularizing means 203

⑥isolated point removing means 205

⑦outline extracting means 206

⑧outline image memory 207

20 ⑨black pixel calculating section 210

⑩comparing section 213

⑪comparator (photograph/character-linear image  
distinguishing section) 215

⑫threshold value

25 ⑬recognition signal

⑭area comparator (headline/drawing distinguishing section) 214

⑮threshold value

⑯text character recognition section 211

⑰rectangular region address generating section 208

5 ⑱black pixel calculating section 209

⑲rectangular area calculating section 212

Fig. 5

20 image input device 25

21 calculation processor 27

10 22 control instruction input device 29

23 display device 26

24 image output device 28

25 memory device 30

Fig. 6

15 26 N

Fig. 7

①image number

②character quantity within the small region in the case of  $k=1$ ,  
 $l=1$

20 Fig. 8

③image number 1

④character information

⑤the entire information of the first image

⑥the entire information of the second image

25 ⑦the entire information of the  $j$ -th image

Fig. 9

- ⑧input desired number  $l$
- ⑨prepare character information  $a_{Kl}$  of image to be retrieved
- ⑩calculate character quantity  $b_{JKl}$  of image number  $j$
- 5 ⑪calculate disagreement  $D_J$  between  $j_{th}$  image and image to be retrieved
- ⑫store number of images from image having the smallest disagreement value to the  $I_{th}$  image
- ⑬process end?

10 Fig. 10

- ①other headline present?
- ②written vertically or horizontally?
- ③horizontal re-arrangement procedure
- ④vertical re-arrangement procedure
- 15 ⑤set upper right end coordinate of character line
- ⑥other headline present?
- ⑦left margin is smaller than right space?
- ⑧execute re-arrangement of character line
- ⑨determine a position subtracted by the width of a character
- 20 from right end of present character line as right end of next character line
- ⑩set upper right end coordinate of drawing and photograph
- ⑪execute re-arrangement of character line
- ⑫change paragraph and set again right end space of character
- 25 line

(a) = 1

(b) = □

Fig. 11

⑬left margin is smaller than right space?

5 ⑭other drawing or photograph present?

⑮execute re-arrangement of drawing or photograph

⑯determine a position subtracted by the width of drawing or photograph from right end of present drawing or photograph as right end of next drawing or photograph

10 ⑰end re-arrangement of drawing or photograph

⑱execute re-arrangement of drawing or photograph

⑲change paragraph and set again right end space of drawing or photograph



⑨ 日本国特許庁(JP)

⑩ 特許出願公開

⑫ 公開特許公報(A) 平3-14184

⑬ Int.Cl.<sup>3</sup>

G 06 F 15/62  
15/40

識別記号

3 2 5 R  
5 3 0 X

庁内整理番号

8125-5B  
7313-5B

⑬ 公開 平成3年(1991)1月22日

審査請求 未請求 請求項の数 1 (全 10 頁)

⑭ 発明の名称 文書画像再配置ファイリング装置

⑮ 特 願 平1-148463

⑯ 出 願 平1(1989)6月13日

⑰ 発 明 者 林 幸 雄 神奈川県海老名市本郷2274番地 富士ゼロックス株式会社  
海老名事業所内

⑱ 出 願 人 富士ゼロックス株式 東京都港区赤坂3丁目3番5号  
社

⑲ 代 理 人 弁理士 岩上 昇一

明 細 書

1 発明の名称

文書画像再配置ファイリング装置

2 特許請求の範囲

文書画像を入力する文書画像入力手段と、

前記文書画像入力手段により入力された文書画像の見出し、本文、図表、写真等の各異なる属性の領域を自動的に抽出、分離する領域分離手段と、

前記領域分離手段により抽出、分離された見出し、本文、図表、写真の各項目を項目ごとに蓄積、格納する項目別ファイル蓄積手段と、

前記項目別ファイル蓄積手段に蓄積されている各項目の内容をそれぞれの項目ごとに検索できる検索手段と、

前記検索手段の検索によって所望の文書画像を検索し、各項目の内容を再配置して全体の文書画像を出力する文書画像再配置手段と、

を備えたことを特徴とする文書画像再配置ファイリング装置。

3 発明の詳細な説明

(産業上の利用分野)

本発明は、新聞切抜等の文書画像において、見出し、本文、図表、写真等異なる属性からなる領域の画像を自動的に抽出し、各項目ごとに蓄積した後、各項目全体を再配置して出力できる文書画像再配置ファイリング装置に関する。

(従来の技術)

新聞切抜等の文書画像をファイリングする装置において、その複雑なレイアウト形状を文書画像の属性(見出し、本文、図表、写真)領域に従って自動的にそれぞれの領域を抽出して、コンパクトに再配置するシステムが提案されている。

たとえば、画像入力装置から入力した新聞画像をマンマシン対話によりパターンとして抽出し、構造記述によりデータを構造化してデータ量を削減して蓄積し、利用時にはその目的に合わせて再編集できるデータベース編集管理システムがある(信学技報 PRL84-101 P. 65~72)。

また、文書の形態的情報やキーワードを用いて、標題、章、節等による階層構造と、本文と図表間の参照構造等の文書構造の自動抽出を行うことができ、さらに、文書構造に基づいた書式とレイアウト知識を用いて、文書のフォーマットや図表の割付けを行う自動レイアウトシステム等がある。(情報処理研究会「文書処理とヒューマンインタフェース」20-3 1988. 9. 8.)。

(発明が解決しようとする課題)

しかし、ファイリングされた文書画像から所望の文書画像を得るためには、文書画像全体を検索・チェックしなければならないので、検索に時間がかかった。また、文書画像には見出し、本文、図表、写真等が混在しているので、データの蓄積として無駄である場合があった。

本発明は文書画像の検索・表示時間を短縮させるとともに、文書画像を再配置することによって文書中の無駄な余白を取り除いてデータ圧縮をさせることができる文書画像再ファイリング装置を

提供することを目的とする。

(課題を解決するための手段)

本発明は、第1図に示すように、新聞切抜等の文書画像をスキャン入力する文書画像入力手段1と、前記文書画像入力手段1により入力された文書画像の見出し(3-1)、本文(3-2)、図表(3-3)、写真(3-4)等の各異なる属性の領域を自動的に抽出、分離する領域分離手段2と、前記領域分離手段2により抽出、分離された見出し、本文、図表、写真等の各項目を項目ごとに蓄積、格納する項目別ファイル蓄積手段3と、前記項目別ファイル蓄積手段3に蓄積されている各項目の内容をそれぞれの項目ごとに検索できる検索手段(6、7)、前記検索手段(6、7)の検索によって所望の文書画像を検索し、各項目の内容を再配置して全体の文書画像を出力する文書画像再配置手段8とから構成する。

(作 用)

文書画像入力手段1によって新聞等をスキャンして入力された入力文書画像は、領域分離処理手

- 3 -

段2で、見出し3-1、本文3-2、図表3-3、写真3-4の各属性領域に分離され、それぞれの領域を囲む矩形の左上端座標(x、y)、幅w、高さhが抽出される。この時、縦書き横書きの判別もなされる。次に、見出し3-1と本文3-2との領域については、文字認識処理が施されてコードデータに変換され、図表3-3と写真3-4との領域については、画像圧縮処理が施された後、各属性項目別にファイリングされる。検索手段6では、見出し3-1あるいは本文3-2に対するキーワード検索(登録済部分キーによる部分一致ストリングマッチング)、または検索手段7では、図表3-3、写真3-4に対する類似画像の検索のいずれかを選んで検索できる。検索時には検索対象とした見出し3-1、本文3-2、図表3-3、写真3-4のいずれかのみ表示し、検索終了時には文書画像全体を再配置処理手段8で再配置処理を施して表示する。再配置処理手段8では、見出し3-1、本文3-2、図表3-3、写真3-4の順に、右上から左に配置し、左端になった

- 4 -

ら段落を変える(横書きの場合は左上から右に配置していく)事をくり返す。

第2図および第3図には各属性領域に分離された結果と再配置された結果の一例をそれぞれ示す。

(実施例)

第1図を参照しつつ本発明の原理を説明する。

図において、文書画像入力手段1では、新聞切抜等の文書画像をスキャン入力し、領域分離手段2で、入力された入力文書画像から文書画像の属性別、たとえば、見出し3-1、本文3-2、図表3-3、写真3-4を自動的に抽出する。この属性項目別に抽出された文書画像から文字として、見出し3-1および本文3-2を文字認識手段4で認識して文字コードに変換した後、項目別ファイル蓄積手段3に蓄積する。また、前記抽出された文書画像が画像である場合には、画像圧縮手段5によって圧縮して項目別ファイル蓄積手段3に蓄積する。蓄積されている文書画像を見たい場合には、文字として見出し3-1および本文3-2を項目別に検索するためのキーワード検索手段6

と、画像として図表 3-3 および写真 3-4 を検索する類似画像検索手段 7 とにより検索できるようになっている。そして、文書画像再配置手段 8 には、これらの検索手段によって、最終検索結果を出力表示する。

本文領域、見出し領域、図表領域、および写真領域のそれぞれの大きさ、領域内の濃度、エッジの数…等の属性の相違に着目することにより、領域の識別および分離を任意に行うことができる。

第 4 図は領域分離手段 2 の構成例を示すブロック図である。これは本出願人の出願になる特願昭 63-252920 号（発明者：中村豊）、特願昭 63-291095 号（発明者：中村豊）、特願平 1-87039 号（発明者：中村豊）等の明細書に開示の技術によって構成した一例であり、2 値化手段 201、矩形化処理された画像を格納する矩形化領域画像メモリ 202、矩形化処理を行う矩形化処理手段 203、2 値化された画像を記憶する 2 値画像メモリ 204、2 値化された画像に対して孤立点除去を施す孤立点除去手段 20

5、孤立点除去後の画像に対して輪郭抽出処理を施す輪郭抽出手段 206、抽出された輪郭画像を記憶する輪郭画像メモリ 207、各画像メモリ 202、204、207 の記憶内容に基づいて画像領域の判定を行う領域識別手段とからなっている。領域判定手段は、矩形領域アドレス発生部 208、黒画素計数部 209、210、本文文字識別部 211、矩形面積算出部 212、比算出部 213、面積比較部（見出し／図表識別部）214、比比較部（写真／文字線画像識別部）215 等からなっている。

入力画像はスキヤナ等より多値で入力される。画像入力データは 2 値化処理手段 201 により 2 値化処理される。2 値化された画像は矩形化処理手段 203 によって黒画素を連結し連結した領域を矩形で囲む矩形化処理を施され、得られた矩形領域は矩形化領域画像メモリ 202 に記憶され、画像の識別単位として利用される。

また、2 値化された画像は、孤立点除去手段 205 により孤立点を除去された後、輪郭抽出手段

- 7 -

206 により輪郭即ちエッジ部分の抽出処理がなされ、輪郭画像メモリ 207 に記憶される。

領域の識別は、矩形化処理手段 203 により得られた矩形で囲まれた領域を単位として、まず、本文文字識別部 211 により、矩形の高さの分布から本文の文字高さを推定して、本文を識別し、これを分離する。次に残りの画像中の見出し領域、図表領域、写真領域を、矩形の大きさ、矩形内の濃度やエッジから判別する。即ち、矩形領域アドレス発生部 208 において、本文文字以外の各矩形領域について矩形化領域画像メモリ 202 から読み出した矩形化処理した画像に基づき矩形領域の各画素のアドレスを発生する。発生したアドレスにより 2 値画像メモリ 204 および輪郭画像メモリ 207 からそれぞれ矩形領域内の画素を読みだし、その読み出した画素の黒画素の数をそれぞれ黒画素計数部 209、210 で計数する。計数した矩形内の 2 値画像の黒画素数と輪郭画像の黒画素数との比を比算出部 213 で算出する。比比較部 215 は比算出部 213 で算出した比を予め

- 8 -

定めたしきい値と比較し、しきい値よりも大きい場合は文字線画像領域、小さい場合は写真領域と判定する。

文字線画像と判定されたときは、矩形の大きさによりそれが見出し領域であるかまたは図表領域であるかを判定する。そのために、矩形化領域画像メモリ 202 から読み出した矩形化処理した画像に基づき矩形領域の面積を矩形面積算出部 212 で算出する。そして面積比較部（見出し／図表識別部）214 において、算出した矩形の面積の大きさを予め定めたしきい値と比較し、その結果、しきい値より小さいときはその矩形領域は見出し領域であると判定（識別）する。比較の結果しきい値よりも大きいときは、図表領域であると判定する。

これらの判定結果（識別信号）に基づいて、イメージメモリ等に記憶されている入力画像から各領域の画像信号が分離抽出され、文字認識や符号圧縮、検索のための属性情報の付加等の必要な処理を施された後、項目別ファイル蓄積部 3 に登録

- 9 -

-541-

- 10 -

される。すなわち、スキャンインされた入力画像は領域分離手段2によって、見出し3-1、本文3-2、図表3-3、写真3-4の各属性領域に分離され、また、それぞれの領域を囲む矩形の左上端座標( $x$ ,  $y$ )、幅 $w$ 、高さ $h$ が抽出される。この時、縦書き横書きの判別もなされる。次に本文および見出し領域については文字認識手段4により文字認識処理が施されてコードデータに変換され、図表、写真領域については画像圧縮手段5により画像圧縮処理が施された後、項目別フェイリング蓄積手段3において各属性項目別にフェイリングされる。

以上のようにして、登録された画像に対する検索は、見出しあるいは本文に対してはキーワード検索手段6によりキーワード検索(登録済部分キーによる部分一致ストリングマッチング)で必要な画像の検索を行う。図表あるいは写真に対しては類似画像検索手段7により行う。

次に、図表3-3および写真3-4に対する類似画像検索の一例について、第5図ないし第9図

を参照しつつ説明する。この例は本出願人の出願になる特願平1-113829号(発明者 広瀬真)の明細書に開示の技術によるものであり、必要とする画像を高速に検索できる利点のあるものである。

第5図は、本発明の上記類似画像検索を行う機能を有する画像検索装置の一実施例の概略の構成を示すブロック図である。本画像検索装置は、利用者あるいは管理者が画像を入力するための画像入力装置25、画像やその他の情報を利用者あるいは管理者に表示するための表示装置26、画像間の一致度の算出や検索された画像の提示順序や装置全体の制御をするための演算装置27、検索された画像を印刷などの形で出力するための画像出力装置28、画像間の一致の判定の際の許容範囲を利用者が入力するための制御入力装置29、画像や情報等を記憶しておく記憶装置30から成っている。

類似画像検索を効率的に行うために、図表および写真を項目別フェイリング蓄積手段3へ登録す

- 1 1 -

る時に加工を施し、特徴情報を付加しておく。

まず、その登録動作を説明する。第8図は、画像入力装置25から送られる二次元二値の画像情報の構造の例を示す図である。この画像情報は表示装置28によって画像として表示される。画像情報全体は $M \times N$ 個の二値情報であり、本実施例では値0は白、1は黒を表すものとする。演算装置27は、 $K \times L$ 個の小領域毎に各領域内の値1の数を量子化した特徴量を第7図の形式にまとめた特徴情報32を作成する。そして、特徴情報32と $M \times N$ 個の画像情報33とを、記憶装置30に記憶する。このようにして記憶装置30には、第5図に示すように多数の画像に対する特徴情報群34と画像情報群35とが記憶される。なお、特徴情報32の先頭の画像番号31は、演算装置27が複数の画像情報を識別するための一意な番号である。

次に画像の検索動作を第9図に示す処理動作フローチャートにより説明する。制御入力装置29からは、一致判定の際の許容範囲として、利用者

- 1 2 -

が必要とする画像の数 $I$ が入力される(①)。

演算装置27は、検索の条件となる画像が与えられると、登録の場合と同様にして検索の条件となる画像に対する特徴情報 $a_{j,x}$ を作成する(②)。

$j$ を1に設定し、不一致度の小さい方から $I$ 番目の値を入れるレジスタ $D_j$ の値を $K \times L$ に設定する(③)。

画像番号 $j$ の画像の特徴量 $b_{j,x}$ を算出する(④)。なお、第8図のように特徴量が登録時に計算され画像のインデックスとして画像と一緒に登録されている場合には、その値を記憶装置30から読み出す。

ここで演算装置27は、画像番号1番から順に $j$ 番目の画像と検索の条件となる画像との間の不一致度 $D_j$ を計算していく(⑤)。

そして $D_j$ が求まる度に、すでに計算してある不一致度の小さい方から1番目の値の $D_1$ と $D_j$ とを比較する(⑥)。

$D_j > D_1$ のときは、1番目までに入っていないので、次の画像についての処理に移るため、 $j$

- 1 3 -

- 542 -

- 1 4 -

の値を1増加する(⑦)。

比較の結果、 $D_j \leq D_i$ のときは、 $D_i$ を $D_j$ の値で置き換える(⑧)。不一致度が小さい方から1番目までの画像の番号を記憶しておく(⑨)。そして、次の画像についての処理に移るため、 $j$ の値を1増加する(⑩)。

最後に、記憶してある全ての画像について不一致度を計算し終ったときに、不一致度の小さい方から1枚の画像を表示装置26に表示する。

ところで、特徴情報群34中における各画像の特徴情報と検索の条件となる画像の特徴情報との間の不一致度を計算する方法を説明する。特徴情報は、画像番号31とK行L列の行列で表せる画像の特徴量から成る。検索の条件となる画像の特徴量を $a_{ki}$  ( $k=1 \sim K$ 、 $i=1 \sim L$ )、特徴情報群34の中のj番目の画像の特徴量を $b_{jki}$

( $k=1 \sim K$ 、 $i=1 \sim L$ )と表す。そして、各小領域毎の領域内の1の数が多いに量子化されているとしたとき、条件となる画像とj番目の画像との不一致度は次式で表される演算によって求め

ることができる。

$$D_j = \sum \sum (a_{ki} \oplus b_{jki})$$

つまり、対応する要素間の排他的論理和の総和を不一致度と定義する。なお $\oplus$ は排他的論理和を表す記号とする。

なお、各小領域毎の領域内の1の数が多いに量子化されているとしたとき、即ち、 $a_{ki}$ と $b_{jki}$ が多値のときは、条件となる画像とj番目の画像との不一致度は次式で表される演算によって求められる。

$$D_j = \sum \sum |a_{ki} - b_{jki}|$$

次に、検索した結果の画像を出力する際には、見出し3-1、本文3-2、図表3-3、写真3-4等文書全体の再配置処理を行う。第10図および第11図は再配置処理動作の手順を示すものである。なお、第10図と第11図とは④と⑤において接続される。

まず、検索された文書画像に対応する蓄積手段3の見出し部の記憶位置を走査して見出し3-1があるか否かを調べる(①)。

- 15 -

見出し3-1がある場合には、その見出し3-1が縦長配置かあるいは横長配置かを調べる(②)。見出し3-1が縦長配置のものであるときは、読み出した見出し文字を縦長配置手順に従って配置を行う(③)。横長配置のものであるときは、読み出した見出し文字を横長配置手順に従って配置する。

見出し3-1が縦長配置(③)および横長配置(④)のいずれの場合でも、さらに他の見出し文字があるか否かを調べる(⑤)。

新たな見出し文字がなければ見出しの配置が終了し、本文の文字列の配置に移る。その際、次の行を文字列の右上端の座標として設定する(⑥)。

文字列がまだあるか否かを調べる(⑦)。文字列があればその文字列の右端からの余白と左側における所定のマージンとを比較して(⑧)、左側マージンの方が小さい間は、文字列を配置する(⑨)。文字列の配置は、現在の文字列の右端から文字幅だけ差し引いた位置が次の文字列の右端となる(⑩)。

- 17 -

- 16 -

処理⑩において、文字列配置の左側余白が、所定の左側マージンより小さくなった場合には、段落の変更と文字列の右端再設定を行う(⑪)。そして、処理⑩と同様に文字列の配置を行う(⑫)。処理⑫で配置すべき文字列がなくなると図表および写真の配置に移るが、まず図表、写真の配置の右上端の座標を設定する(⑬)。

処理⑬で検索結果の文書画像に図表または写真があれば、それを配置すべき右端からの左側余白と左側における所定のマージンとを比較して(⑭)、左側マージンの方が小さい間は、図表または写真を配置する(⑮)。図表または写真の配置は、現在の図表または写真の右端から図表または写真の幅だけ差し引いた位置が次の図表または写真の右端となる(⑯)。

処理⑯において、左側マージンの方が大きくなった場合には、段落の変更と図表または写真の右端再設定を行う(⑰)。そして、処理⑯と同様に図表または写真の配置を行う(⑱)。処理⑱で図表または写真がなくなると再配置は終了する(

- 543 -

- 18 -

の)。

(発明の効果)

本発明によれば、文書画像における見出し、本文、図表、写真等異なる属性の領域を自動的に抽出、分離し、項目(属性領域)ごとに分けて蓄積するとともに、検索においては属性の領域ごとに適した検索方法により検索できるようにし、検索時の表示はその属性の領域のみとすることができ、検索を効率的に行うことができる。

さらに、検索結果の文書画像を全体を項目ごとにそれぞれまとめて再配置するようにしたので、文書中の無駄な余白を取り除きデータを圧縮させることができ、また表示に要する時間を短縮させることができる。

#### 4 図面の簡単な説明

第1図は本発明の文書画像再配置ファインリング装置の構成を示す図、第2図および第3図は領域分離で抽象された各属性を項目別にファイルする様子を示した図、第4図は領域分離手段のブロック構成図、第5図は画像検索装置のブロック構成

図、第6図は二次元二値の画像情報の構造を示す図、第7図は特徴情報の構造を示す図、第8図は特徴情報群と画像情報群とを示す図、第9図は検索動作時の処理フローチャート、第10図および第11図は第1図図示における文書画像再配置処理手段8のフローチャートを示す。

- 1・・・文字画像入力手段
- 2・・・領域分離手段
- 3・・・項目別ファイル蓄積手段
- 4・・・文字認識手段
- 5・・・画像圧縮手段
- 6・・・キーワード検索手段
- 7・・・類似画像検索手段
- 8・・・文書画像再配置処理手段

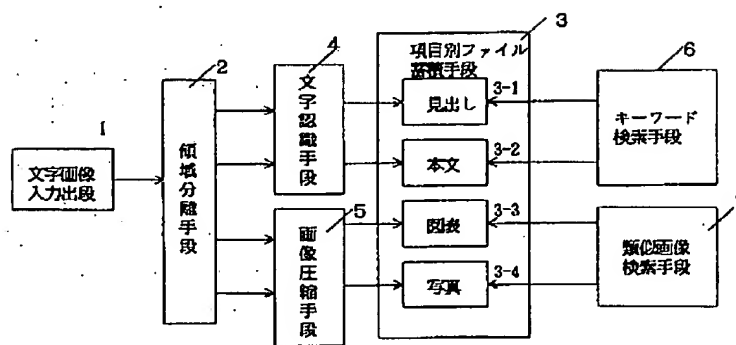
特許出願人 富士ゼロックス株式会社  
代理人 弁理士 岩 上 昇



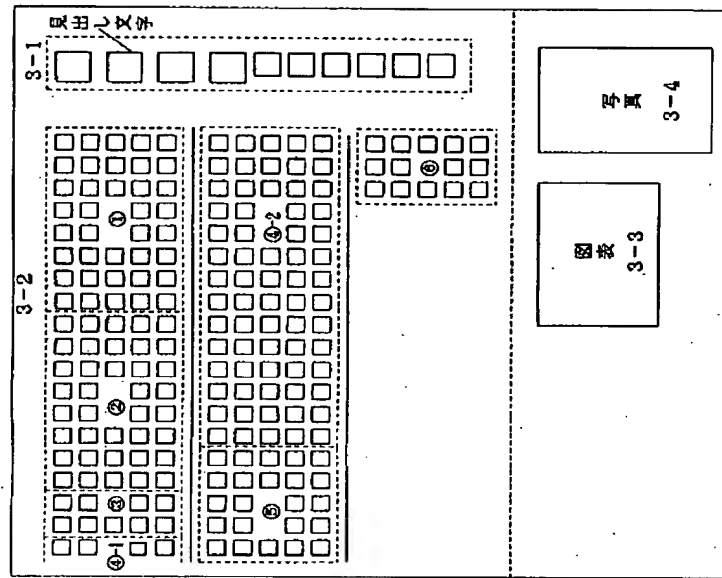
- 19 -

- 20 -

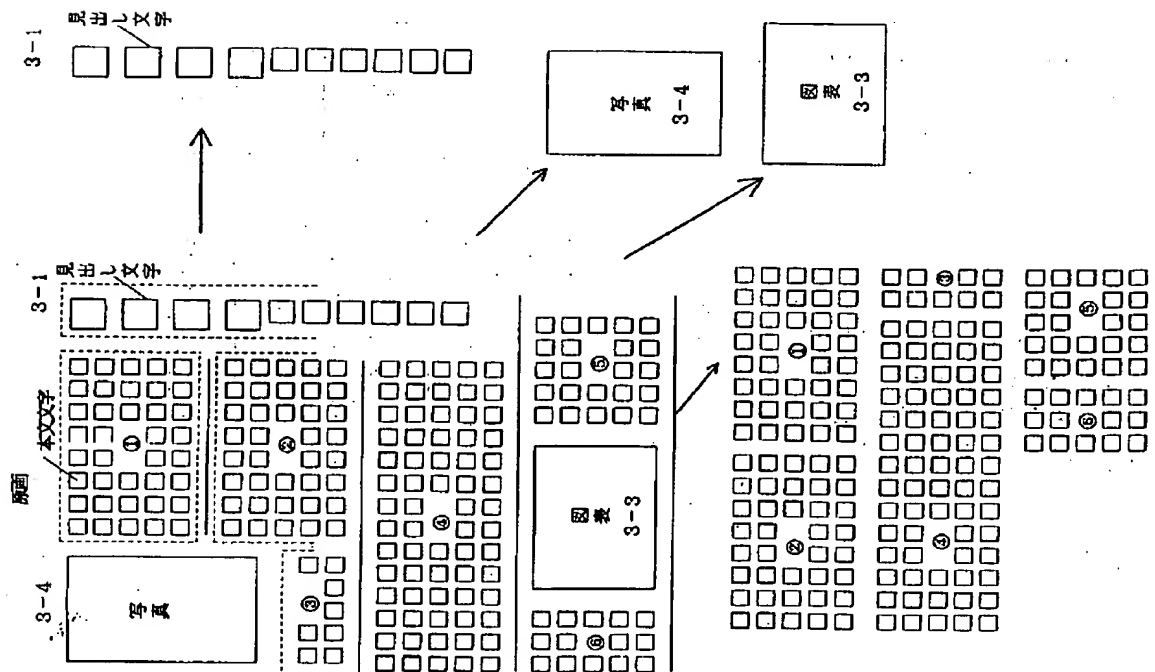
第1図



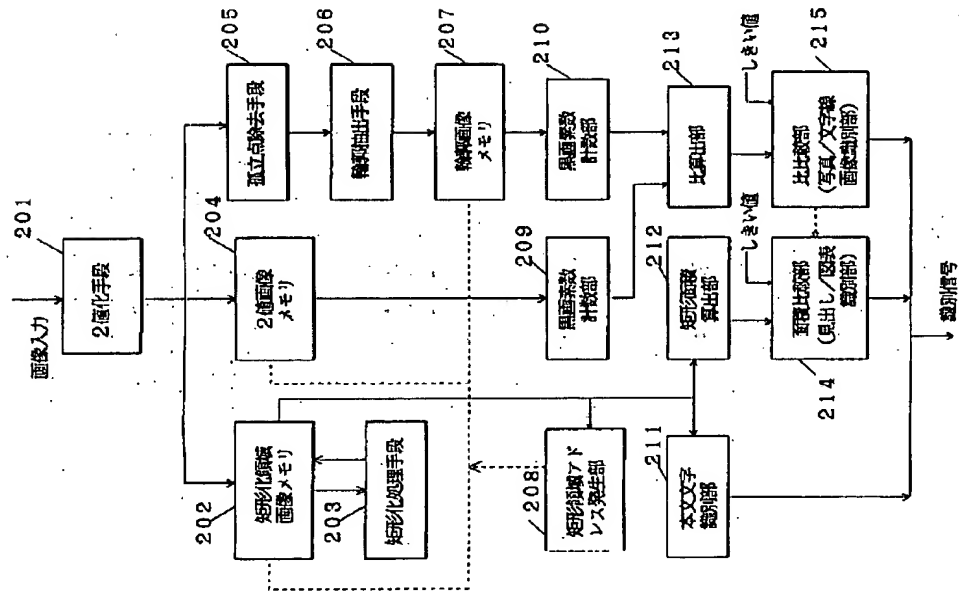
第3図



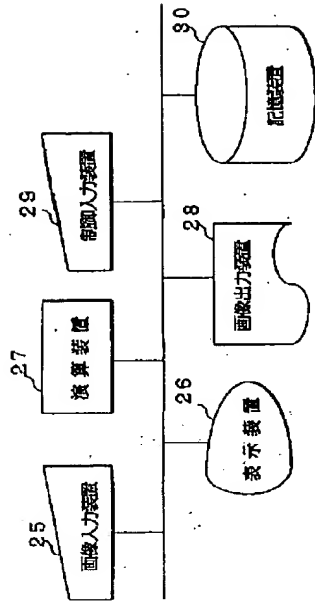
第2図



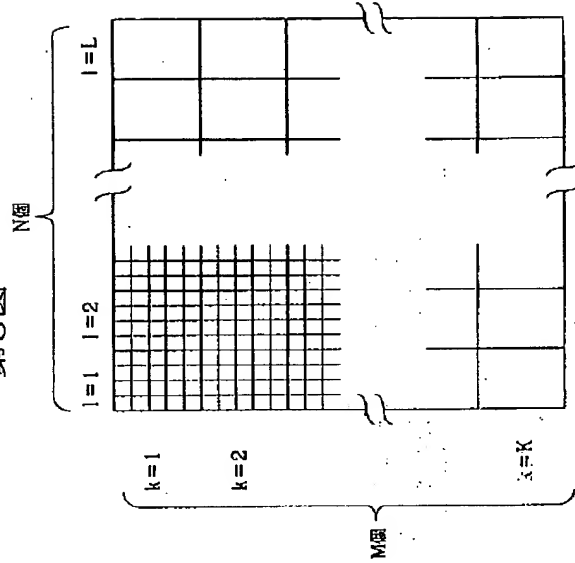
第4図



第5図

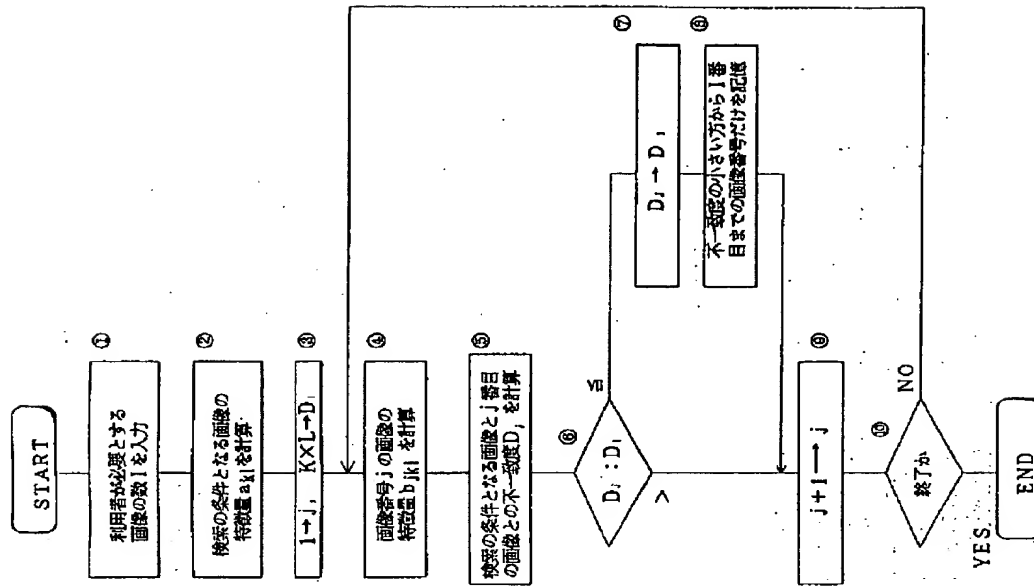


第6図

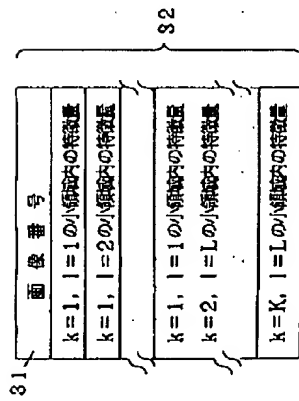




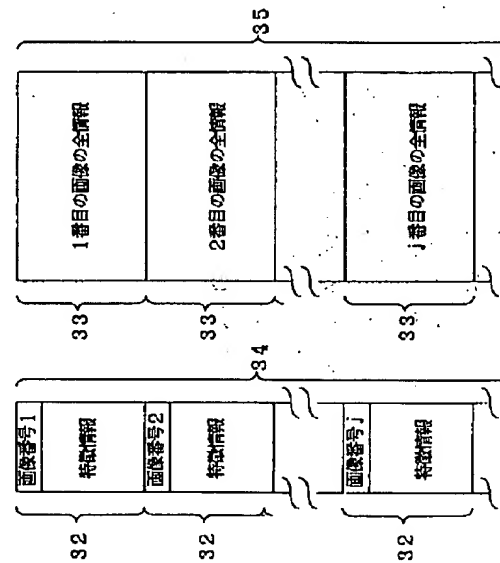
第9図



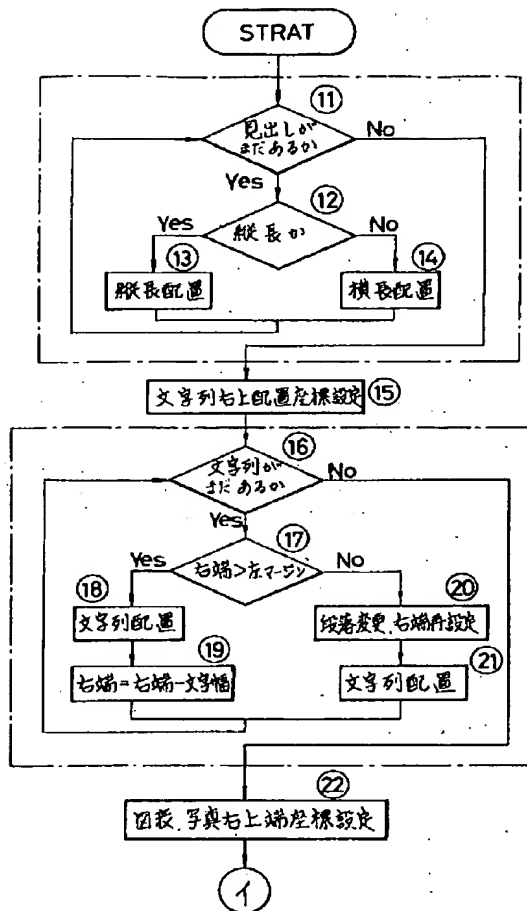
第7図



第8図



第 10 図



第 11 図

